(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平11－102372

(43)公開日　平成11年(1999)４月13日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３２０Ａ
17/27		15/20	５５０Ｆ
		15/40	３７０Ａ

審査請求　未請求　請求項の数８　ＯＬ　（全９頁）

(21)出願番号　特願平9－263318

(22)出願日　平成９年(1997)９月29日

(71)出願人　000005049
シャープ株式会社
大阪府大阪市阿倍野区長池町22番22号

(72)発明者　池内　洋
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内

(72)発明者　芥子　育雄
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内

(72)発明者　黒武者　健一
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内

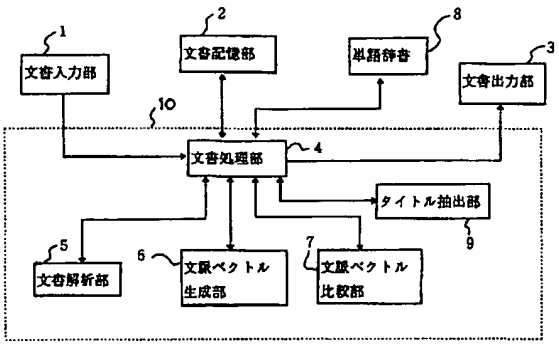
(74)代理人　弁理士　梅田　勝

(54)【発明の名称】　文書要約装置及びコンピュータ読み取り可能な記録媒体

(57)【要約】

【課題】　新聞記事から要約を抽出するには、文書の意味を扱う必要があるが、従来は特定の文書形式や文脈を仮定する必要があった。

【解決手段】　タイトル抽出部９は文書入力部１から入力された文書からタイトルと本文を抽出し、文書解析部５はタイトルを単語に、本文を文および単語に分解する。文脈ベクトル生成部６は、単語辞書８を用いてタイトルおよび本文中の文の文脈ベクトルを生成する。文脈ベクトル比較部７は、タイトルと本文中の各文の文脈ベクトルを比較して各文脈ベクトル間距離を算出する。文書処理部４は各文脈ベクトル間距離を参照して、タイトルに近い複数文を生成して文書の重要部分とする。このように、入力文書を文脈ベクトルを用いて解析することによって、特定の文書形式や文脈を仮定することなく、質の良い重要部分を簡単な処理で抽出できる。



1

【特許請求の範囲】

【請求項１】　入力された文書のタイトルと本文から要約を作成する文書要約装置であって、
単語の文脈ベクトルが格納された単語辞書と、
上記タイトルを単語に、上記本文を文および単語にそれぞれ分解する文書解析手段と、
上記単語辞書を参照して上記タイトルの文脈ベクトル及び上記分割された文の文脈ベクトルを生成する文脈ベクトル生成手段と、
上記タイトルの文脈ベクトルと上記分割された文の文脈ベクトルとの距離を算出する距離算出手段と、
上記算出された距離に基づいて上記本文から少なくとも一つの文を要約として選出する選択手段と、を備えることを特徴とする文書要約装置。

【請求項２】　入力された文書のタイトルと本文から要約を作成する文書要約装置として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、
単語の文脈ベクトルが格納された単語辞書と、
上記タイトルを単語に、上記本文を文および単語にそれぞれ分解する文書解析手段と、
上記単語辞書を参照して上記タイトルの文脈ベクトル及び上記分割された文の文脈ベクトルを生成する文脈ベクトル生成手段と、
上記タイトルの文脈ベクトルと上記分割された文の文脈ベクトルとの距離を算出する距離算出手段と、
上記算出された距離に基づいて上記本文から少なくとも一つの文を要約として選出する選択手段として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項３】　入力された文書のタイトルと本文から要約を作成する文書要約装置であって、
単語の文脈ベクトルが格納された単語辞書と、
上記タイトルを単語に、上記本文を文および単語にそれぞれ分解する文書解析手段と、
上記単語辞書を参照して上記タイトルの文脈ベクトル及び上記分割された文の文脈ベクトル及び上記文書全体の文脈ベクトルを生成する文脈ベクトル生成手段と、
上記タイトルの文脈ベクトルと上記分割された文の文脈ベクトルとの第１の距離を算出すると共に、上記文書全体の文脈ベクトルと上記分割された文の文脈ベクトルとの第２の距離を算出する距離算出手段と、
上記算出された第１の距離に基づいて上記本文から少なくとも一つの文を要約として選出すると共に、上記算出された第２の距離に基づいて上記本文から少なくとも一つの文を要約として選出する選択手段と、を備えることを特徴とする文書要約装置。

【請求項４】　入力された文書のタイトルと本文から要約を作成する文書要約装置として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体

2

であって、
単語の文脈ベクトルが格納された単語辞書と、
上記タイトルを単語に、上記本文を文および単語にそれぞれ分解する文書解析手段と、
上記単語辞書を参照して上記タイトルの文脈ベクトル及び上記分割された文の文脈ベクトル及び上記文書全体の文脈ベクトルを生成する文脈ベクトル生成手段と、
上記タイトルの文脈ベクトルと上記分割された文の文脈ベクトルとの第１の距離を算出すると共に、上記文書全体の文脈ベクトルと上記分割された文の文脈ベクトルとの第２の距離を算出する距離算出手段と、
上記算出された第１の距離に基づいて上記本文から少なくとも一つの文を要約として選出すると共に、上記算出された第２の距離に基づいて上記本文から少なくとも一つの文を要約として選出する選択手段として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項５】　入力された文書のタイトルと本文から要約を作成する文書要約装置であって、
単語の文脈ベクトルが格納された単語辞書と、
上記タイトルを単語に、上記本文を文および単語にそれぞれ分解する文書解析手段と、
上記単語辞書を参照して上記タイトルの文脈ベクトル及び上記分割された文の文脈ベクトル及び上記文書全体の文脈ベクトルを生成する文脈ベクトル生成手段と、
上記タイトルの文脈ベクトルと上記分割された文の文脈ベクトルとの第１の距離を算出すると共に、上記文書全体の文脈ベクトルと上記分割された文の文脈ベクトルとの第２の距離を算出する距離算出手段と、
上記算出された第１の距離と第２の距離の和に基づいて上記本文から少なくとも一つの文を要約として選出する選択手段と、を備えることを特徴とする文書要約装置。

【請求項６】　入力された文書のタイトルと本文から要約を作成する文書要約装置として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、
単語の文脈ベクトルが格納された単語辞書と、
上記タイトルを単語に、上記本文を文および単語にそれぞれ分解する文書解析手段と、
上記単語辞書を参照して上記タイトルの文脈ベクトル及び上記分割された文の文脈ベクトル及び上記文書全体の文脈ベクトルを生成する文脈ベクトル生成手段と、
上記タイトルの文脈ベクトルと上記分割された文の文脈ベクトルとの第１の距離を算出すると共に、上記文書全体の文脈ベクトルと上記分割された文の文脈ベクトルとの第２の距離を算出する距離算出手段と、
上記算出された第１の距離と第２の距離の和に基づいて上記本文から少なくとも一つの文を要約として選出する選択手段として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項７】　上記距離算出手段が、上記タイトルと上記分割された文との同一単語の比率を考慮して第１の距離を算出するとともに、上記文書全体と上記分割された文との同一単語の比率を考慮して第２の距離を算出することを特徴とする請求項３または請求項５に記載の文書要約装置。

【請求項８】　上記単語辞書が、単語の特徴ベクトルと共に単語の重要度を格納し、上記距離算出手段が、上記同一単語の比率を算出する際に、上記単語の重要度を用いることを特徴とする請求項７に記載の文書要約装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、電子新聞や情報検索装置等で用いられ、情報の内容を要約して利用者に提示する文書要約装置に関するものである。

【０００２】

【従来の技術】電子新聞や情報検索装置は、利用者に情報を提示する装置である。電子新聞は各新聞記事を項目として利用者に提示し、利用者は提示された項目から、自分にとって必要と思われるものを選び、その記事内容を読むことによって情報を得るものである。情報検索装置は、利用者が与えた検索要求に基づいてデータを検索し、検索された各データを項目として利用者に提示し、利用者は電子新聞と同様に、提示された項目から自分にとって必要と思われるものを選び、そのデータ内容を読むことによって情報を得るものである。

【０００３】特に新聞記事は、その内容は一般に複数の文からなり、利用者はそのすべてを読むことで、内容を完全に理解することが出来る。従って利用者にとって、記事の文章量が多くなるほど、内容を理解するのに時間と労力が必要とされる。そこで、従来では、利用者の読むべき文章量を減少させるように文章中から要約を抽出することがなされていた。

【０００４】ここで、個々の文書の文章量を減少させることによって利用者の手間を軽減するために、文書から要約を抽出する場合を考える。この場合には、文書の文章量を減少させても元の文書の含まれる重要な内容が損なわれないような手法を用いる必要がある。

【０００５】従来から提唱されている文書要約の手法としては、主に次の２つの手法がある。

【０００６】第１の手法は、文章を表層的に解析するものである。この手法は、単語の出現頻度解析から文章の重要箇所を決定して元の文書に含まれている単語の組み合わせや文の抽出によって要約文の生成を行うものや、文の文末表現および用語によって文章中における主張文を抽出するものがある。

【０００７】第２の手法は、文章を意味的に解析するものである。この手法は、事前に文章の形式や文脈を仮定しておいて、その仮定に沿って文章を解析して要約を抽出するものや、文の係り受けの粗密性を用いることによって内容の重要性を定義して要約を抽出するものがある。

【０００８】また、上述とは全く異なる第３の手法として、特開平６－２１５０４９号公報に示されるように、文脈ベクトルによって、全体の文章と最も意味の近い段落あるいは各段落に最も意味の近い文を求め、それらを要約として提示するものである。

【０００９】

【発明が解決しようとする課題】ところが、第１の手法と第２の手法には各々以下のような問題がある。すなわち、第１の手法は、第２の手法に比べて簡単に実施できる反面、意味を扱わないので文書中の不要な箇所を重要な箇所と誤って判断してしまうという問題がある。一方、第２の手法は、最初の仮定が当てはまらないようなタイプの文書に対しては全く非力であり、内容の重要性の定義自体が困難である上、第１の手法に比べて処理が複雑であるという問題がある。

【００１０】第３の手法は、特定の文書形式や文脈を仮定することなく、簡単な処理によって文書における重要部分を要約として抽出できるが、文書全体あるいは段落全体の文脈ベクトルが、要約を最も良く表すとは限らない。特に新聞記事においては、記事全体よりもタイトルに要約がより良く表現されていると考えられる。

【００１１】本発明の目的は、新聞記事のように、タイトルとその本文からより精度の高い要約を求めることのできる文書要約装置を提供することにある。

【００１２】

【課題を解決するための手段】請求項１に記載の文書要約装置は、入力された文書のタイトルと本文から要約を作成する文書要約装置であって、単語の文脈ベクトルが格納された単語辞書と、上記タイトルを単語に、上記本文を文および単語にそれぞれ分解する文書解析手段と、上記単語辞書を参照して上記タイトルの文脈ベクトル及び上記分割された文の文脈ベクトルを生成する文脈ベクトル生成手段と、上記タイトルの文脈ベクトルと上記分割された文の文脈ベクトルとの距離を算出する距離算出手段と、上記算出された距離に基づいて上記本文から少なくとも一つの文を要約として選出する選択手段と、を備えることを特徴とする。

【００１３】請求項２に記載のコンピュータ読み取り可能な記録媒体は、入力された文書のタイトルと本文から要約を作成する文書要約装置として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、単語の文脈ベクトルが格納された単語辞書と、上記タイトルを単語に、上記本文を文および単語にそれぞれ分解する文書解析手段と、上記単語辞書を参照して上記タイトルの文脈ベクトル及び上記分割された文の文脈ベクトルを生成する文脈ベクトル生成手段と、上記タイトルの文脈ベクトルと上記分割された文の文脈ベクトルとの距離を算出する距離算出手段と、上記算出さ

5

れた距離に基づいて上記本文から少なくとも一つの文を要約として選出する選択手段として機能させるためのプログラムを記録している。

【００１４】請求項３に記載の文書要約装置は、入力された文書のタイトルと本文から要約を作成する文書要約装置であって、単語の文脈ベクトルが格納された単語辞書と、上記タイトルを単語に、上記本文を文および単語にそれぞれ分解する文書解析手段と、上記単語辞書を参照して上記タイトルの文脈ベクトル及び上記分割された文の文脈ベクトル及び上記文書全体の文脈ベクトルを生成する文脈ベクトル生成手段と、上記タイトルの文脈ベクトルと上記分割された文の文脈ベクトルとの第１の距離を算出すると共に、上記文書全体の文脈ベクトルと上記分割された文の文脈ベクトルとの第２の距離を算出する距離算出手段と、上記算出された第１の距離に基づいて上記本文から少なくとも一つの文を要約として選出すると共に、上記算出された第２の距離に基づいて上記本文から少なくとも一つの文を要約として選出する選択手段と、を備えることを特徴とする。

【００１５】請求項４に記載のコンピュータ読み取り可能な記録媒体は、入力された文書のタイトルと本文から要約を作成する文書要約装置として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、単語の文脈ベクトルが格納された単語辞書と、上記タイトルを単語に、上記本文を文および単語にそれぞれ分解する文書解析手段と、上記単語辞書を参照して上記タイトルの文脈ベクトル及び上記分割された文の文脈ベクトル及び上記文書全体の文脈ベクトルを生成する文脈ベクトル生成手段と、上記タイトルの文脈ベクトルと上記分割された文の文脈ベクトルとの第１の距離を算出すると共に、上記文書全体の文脈ベクトルと上記分割された文の文脈ベクトルとの第２の距離を算出する距離算出手段と、上記算出された第１の距離に基づいて上記本文から少なくとも一つの文を要約として選出すると共に、上記算出された第２の距離に基づいて上記本文から少なくとも一つの文を要約として選出する選択手段として機能させるためのプログラムを記録している。

【００１６】請求項５に記載の文書要約装置は、入力された文書のタイトルと本文から要約を作成する文書要約装置であって、単語の文脈ベクトルが格納された単語辞書と、上記タイトルを単語に、上記本文を文および単語にそれぞれ分解する文書解析手段と、上記単語辞書を参照して上記タイトルの文脈ベクトル及び上記分割された文の文脈ベクトル及び上記文書全体の文脈ベクトルを生成する文脈ベクトル生成手段と、上記タイトルの文脈ベクトルと上記分割された文の文脈ベクトルとの第１の距離を算出すると共に、上記文書全体の文脈ベクトルと上記分割された文の文脈ベクトルとの第２の距離を算出する距離算出手段と、上記算出された第１の距離と第２の距離の和に基づいて上記本文から少なくとも一つの文を

6

要約として選出する選択手段と、を備えることを特徴とする。

【００１７】請求項６に記載のコンピュータ読み取り可能な記録媒体は、入力された文書のタイトルと本文から要約を作成する文書要約装置として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、単語の文脈ベクトルが格納された単語辞書と、上記タイトルを単語に、上記本文を文および単語にそれぞれ分解する文書解析手段と、上記単語辞書を参照して上記タイトルの文脈ベクトル及び上記分割された文の文脈ベクトル及び上記文書全体の文脈ベクトルを生成する文脈ベクトル生成手段と、上記タイトルの文脈ベクトルと上記分割された文の文脈ベクトルとの第１の距離を算出すると共に、上記文書全体の文脈ベクトルと上記分割された文の文脈ベクトルとの第２の距離を算出する距離算出手段と、上記算出された第１の距離と第２の距離の和に基づいて上記本文から少なくとも一つの文を要約として選出する選択手段として機能させるためのプログラムを記録している。

【００１８】請求項７に記載の文書要約装置は、請求項３または請求項５に記載の文書要約装置において、上記距離算出手段が、上記タイトルと上記分割された文との同一単語の比率を考慮して第１の距離を算出するとともに、上記文書全体と上記分割された文との同一単語の比率を考慮して第２の距離を算出することを特徴とする。

【００１９】請求項８に記載の文書要約装置は、請求項７に記載の文書要約装置において、上記単語辞書が、単語の特徴ベクトルと共に単語の重要度を格納し、上記距離算出手段が、上記同一単語の比率を算出する際に、上記単語の重要度を用いることを特徴とする。

【００２０】

【発明の実施の形態】

（実施の形態１）以下、本実施の形態を図面を用いて説明する。図１に本実施の形態の文書要約装置におけるブロック図を示す。要約作成の対象となる文書が入力される文書入力部１は、キーボードや光学式文字読み取り装置（ＯＣＲ）、あるいは着脱式外部記憶装置で構成され、処理部１０に接続されている。この接続は、通信回線を介して接続されていてもよい。文書記憶部２は、文書入力部１から入力された文書、及び処理部１０で生成された要約が格納される。入力文書や要約を出力する文書出力部３は、ＣＲＴ（カソード・レイ・チューブ）、液晶表示装置（ＬＣＤ）、プリンタ、あるいは着脱式外部記憶装置で構成されている。この接続は、処理部１０に接続され、通信回線を介して接続されていてもよい。

【００２１】処理部１０内の文書処理部４は、編集や検索等の一般的な文書処理を実施する他に、以下に述べる文書解析部５、文脈ベクトル生成部６、文脈ベクトル比較部７およびタイトル抽出部９を制御して、入力文書の要約を生成する。

【００２２】まず最初に、本発明で用いる特徴ベクトルとしての文脈ベクトルについて簡単に説明する。特徴語として、例えば次のような複数の単語「人間，男，女，機械，知識，活動，経験，政治，芸術，科学，…」を用意して特徴空間を定義する。上記特徴語の個数は任意であるが、少なくとも２００語～５００語程度は用意しておく方が実用上は望ましい。また、特徴語の種類や分野も任意であり、選択に当たっての厳密さは要求されず、特徴が相互にオーバラップしていても構わない。さらに、要約抽出の対象となる文書の分野が特定の分野である場合には、その分野に特有の特徴語を予め充実させることによって、より品質の高い要約を抽出できることになる。

【００２３】単語辞書８には、文脈ベクトルを生成する際に使用される単語を格納し、単語辞書８に格納された各単語と上記特徴語との関連の有無に応じて当該単語を上記特徴空間に配置する。その際における各単語の特徴空間上の位置がその単語の文脈ベクトルであり、この文脈ベクトルは単語に対応させて単語辞書８に格納されている。

【００２４】図２に、各単語の文脈ベクトルが定義された単語辞書８の内容の一例を示す。ここで、各要素の配列順序は上述した特徴語の配列順序と同じである。各単語の文脈ベクトルは、単語辞書８内に格納されている単語と各特徴語との関連をその有無によって表現した数字を要素とするベクトルである。すなわち、図２において、関連がある場合には要素“１”を与え、関連がない場合には要素“０”を与えている。ここでは、“１”と“０”を与えているが、これに限定されるものではなく、関連の強度を段階的に表現した数字を要素として与えてもよい。

【００２５】図２に例示された単語の文脈ベクトルは以下のことを表現している。すなわち、「人間」という単語は、各特徴語“人間”，“男”，“女”，…とは関連があり、各特徴語“機械”，“知識”，“活動”，“経験”，“政治”，“芸術”，“科学”，…とは関連がないという特徴を表現している。また、「自動車」という単語は、各特徴語“人間”，“男”，“女”，“知識”，“経験”，“政治”，“芸術”，“科学”，…とは関連がなく、各特徴語“機械”，“活動”，…とは関連があるという特徴を表現しているのである。

【００２６】本実施の形態において文脈ベクトルを生成する際に用いる単語は、“名詞”および“サ変名詞（語尾に「する」と付けるとサ行変格活用動詞になる名詞）”だけである。したがって、単語辞書８に登録されている単語も名詞およびサ変名詞である。なお、単語辞書８に登録されている単語は名詞およびサ変名詞にしているが、これに限定されるものではない。

【００２７】次に、本実施の形態の要約作成処理について図３と図４を用いて説明する。図３は、文書処理部４による要約作成処理動作のフローチャートである。図４は、電子化されたタイトルとその本文からなるニュース記事である。

【００２８】まず、ステップＳ１で、文書入力部１から要約抽出の対象となる文書が入力されて図４に示す内容すべてが文書記憶部２に記憶される。ステップＳ２で、タイトル抽出部９によってタイトルと本文が抽出される。タイトル抽出部９は、文書記憶部２から読み出した文書を解析してタイトルと本文とを区別して抽出する。なお、予めタイトルと本文が区別されて入力される場合には、このステップは不要である。一般に電子化されたニュース記事では、予め記事本文とタイトル、およびそれに付随する記事の作成日時、ジャンルなどの情報が一定のフォーマットで付加されている。その場合には単にタイトルに対応する項目と本文に対応する項目から、タイトルと本文とを区別できる。また、タイトルは一般に文の先頭に位置している、あるいはフォントの大きい文字が使われている等のよく知られている特徴を使ってタイトルを抽出し、全文からタイトルを除いたものを本文とするような方法で区別してもよい。以上の方法で抽出されたタイトルと本文は、文書記憶部２の中では区別されて記憶されている。

【００２９】ステップＳ３で、文書解析部５によって、文書記憶部２から本文が読み出されて文単位に分割される。その際に、例えば句点を文の区切りとする。

【００３０】ステップＳ４で、文書解析部５によって、文書記憶部２からタイトルおよび本文が読み出され、この読み出されたタイトルと本文が形態素解析（文章を文法的に解析して単語に分割し、各単語の品詞や活用形等の情報を抽出する手法）によって単語に分解される。図４の例では「ノンバンク整理手続き」がタイトルであり、それが「ノンバンク」「整理」「手続き」の３単語に分解される。文の区切りを句点にして、本文は「日本債券信用銀行の・・・停止した」「一日にも不良債券処理の・・・倒産する」等の文に分割され、前者はさらに「日本」「債券」「信用」「銀行」「の」等の単語に分割され、後者は「一日」「に」「も」「不良」「債券」「処理」「の」等の単語に分解される。このようにして、文書解析部５によってタイトルを単語に、本文を文および単語に分解する。なお、本実施の形態では、単語を抽出する際に形態素解析を行っているが、単語辞書８にある単語との文字列のマッチングによって単語を抽出する方法を用いてもよい。

【００３１】そして、得られた単語のうち名詞およびサ変名詞のみが、タイトルおよび上記ステップＳ３において分割された各文の単位で、文脈ベクトル生成部６に送出される。

【００３２】ステップＳ５で、文脈ベクトル生成部６によって、タイトルの文脈ベクトルと各文の文脈ベクトルとが次のようにして生成される。上記タイトルを構成す

る単語、及び各文を構成する単語の文脈ベクトルは、単語辞書８を引くことによって得られ、さらに、タイトルを構成する単語の文脈ベクトルが加算され正規化されてタイトルの文脈ベクトルが得られ、同様に各文を構成する単語の文脈ベクトルが加算され正規化されて各文の文脈ベクトルが得られるのである。ここで、文脈ベクトルの正規化とは、文脈ベクトルの長さを一定の値に揃えることである。

【００３３】この処理を図４を用いて具体的に説明する。タイトルを構成する単語は「ノンバンク」「整理」「手続き」で、それらは名詞またはサ変名詞である。よって、タイトルの文脈ベクトルは、それら３単語の文脈ベクトルを単語辞書８から取り出して加算し正規化して得られる。具体的な計算例として、例えば特徴語が１２個で、３単語の文脈ベクトルが（0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0），（0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0）（0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0）の時、正規化の際に長さを１０に揃えるとすると、まずそれらを加算したものは、（0, 0, 0, 1, 3, 3, 2, 2, 0, 0, 0, 0）で、その長さは $(0^2+0^2+0^2+1^2+3^2+\cdots+0^2)^{0.5}=27^{0.5}$ である。よって、それを正規化したタイトルの文脈ベクトルは、$10/27^{0.5}\times$（0, 0, 0, 1, 3, 3, 2, 2, 0, 0, 0, 0）≒（0.0, 0.0, 0.0, 1.9, 5.8, 5.8, 3.8, 3.8, 0.0, 0.0, 0.0, 0.0）と求められる。本文についても全く同様にして各文の文脈ベクトルが求められる。

【００３４】ステップＳ６で、文脈ベクトル比較部７によって、上記ステップＳ５において得られたタイトルの文脈ベクトルと各文の文脈タイトルとが比較されて各文脈ベクトル間の距離が算出される。その際に算出される２つの文脈ベクトル間の距離は、正規化された当該両文脈ベクトルの内積を用いる。そして、内積値が大きいほど距離が近いのである。つまり、タイトルとの意味が近く、要約としてふさわしい文である。

【００３５】ステップＳ７で、文書処理部４によって、ステップＳ６において算出されたタイトルと各文との文脈ベクトル間の距離が参照されて、この距離の近い順に本文中の各文が所定数だけ入力文書の要約文として文書記憶部２に格納され、必要に応じて文書出力部３から出力される。文書出力部３は、この要約文のみを出力したり、要約文の箇所をアンダーラインや反転等によって強調された文書全体を出力したりしてユーザの便宜を図ることができる。

【００３６】このステップＳ６とステップＳ７の処理を具体的に説明する。上述の結果からタイトルの文脈ベクトルは、

（0.0, 0.0, 0.0, 1.9, 5.8, 5.8, 3.8, 3.8, 0.0, 0.0, 0.0, 0.0）であり、

ステップＳ５で求められた、本文中の最初の２文の文脈ベクトルが

（0.0, 0.0, 0.0, 6.4, 2.1, 2.1, 6.4, 2.1, 0.0, 0.0, 2.1, 0.0）

（0.0, 4.2, 0.0, 0.0, 6.3, 2.1, 0.0, 0.0, 0.0, 6.3, 0.0, 0.0）

であるとする。タイトルの文脈ベクトルと本文中の最初の文の文脈ベクトルとの内積は、

0.0×0.0+0.0×0.0+0.0×0.0+1.9×6.4+5.8×2.1+5.8×2.1+3.8×6.4+3.8×2.1+0.0×0.0+0.0×0.0+0.0×2.1+0.0×0.0＝68.82

タイトルの文脈ベクトルと本文中の２番目の文の文脈ベクトルとの内積は、

0.0×0.0+0.0×4.2+0.0×0.0+1.9×0.0+5.8×6.3+5.8×2.1+3.8×0.0+3.8×0.0+0.0×0.0+0.0×6.3+0.0×0.0+0.0×0.0＝48.72となる。

【００３７】よって、この場合本文中の最初の文の方が２番目の文よりタイトルに近いと判断する。このような内積計算を本文中の全ての文について行い、内積の大きい順に所定数の文が出力されるのである。

【００３８】（実施の形態２）本実施の形態は、実施の形態１と従来の第３の手法とを組合わせて、双方から得られた要約文を共に出力することで、より要約の精度を向上させるものである。

【００３９】本実施の形態では、実施の形態１の要約処理に加え、単語辞書８を参照してタイトルと本文の文書全体に含まれる単語から文書全体の意味を表す文脈ベクトルを実施の形態１と同様にして生成し、文脈ベクトル比較部７を用いて文書全体の文脈ベクトルと各文の文脈ベクトルの内積計算を行って距離を求め、距離の近い文ほど文書全体の意味に近いものとなる。双方の手法により求めた距離の順に、タイトルとの意味が近い要約文と、文書全体との意味の近い要約文を共に出力して入力文書の要約となる。

【００４０】（実施の形態３）本実施の形態は、実施の形態２において双方の手法により求めた特徴ベクトル間の距離から総合距離を求めて、総合距離の順に得られた要約文を出力することで、より要約の精度を向上させるものである。

【００４１】本実施の形態では、タイトルの文脈ベクトルと各文の文脈ベクトルとの内積値と文書全体の文脈ベクトルと各文の文脈ベクトルとの内積値とを用いることで実現できる。要約を行う文書の各文に文番号を付け、文番号ｉの文脈ベクトルとタイトルの文脈ベクトルとの内積値をＳＴ（ｉ）、文番号ｉの文脈ベクトルと文書全体の文脈ベクトルとの内積値をＳＡ（ｉ）とすると、実施の形態２の２つの方法で計算される距離を総合した総合距離Ｓ（ｉ）は、

Ｓ（ｉ）＝ｆ（ＳＴ（ｉ），ＳＡ（ｉ））

となる。ここで、ｆは何らかの関数であり、単純に加算もしくはそれらの加重和を取る。具体例としては、

Ｓ（ｉ）＝α×ＳＴ（ｉ）＋ＳＡ（ｉ）

となる。ここで、αは定数である。このようにして求め

た距離ＳＴ（ｉ）の順に、要約文が出力され、入力文書の要約となる。αの設定は、文書出力部３の出力結果を見ながらユーザが文書入力部１から設定できるようにしてもよく、また、出力された要約文の所定順位までのＳＴ（ｉ）の総和とＳＡ（ｉ）の総和とが均等になるように文書処理部４で自動設定するようにしてもよい。このようにすることで、双方の手法による要約が必ず含まれることになる。なお、α＝０の時は、文書全体の文脈ベクトルとの内積を取った場合と同じになり、α≫１の時はタイトルの文脈ベクトルとの内積値だけで要約文を選ぶ場合と同じになる。

【００４２】（実施の形態４）本実施の形態は、実施の形態１の要約処理に加え、比較する文に含まれている単語の一致する割合を考慮した従来の第１の手法を取り入れて、タイトルと各文の距離や文書全体と各文の距離を計算して要約文を出力することで、より要約の精度を向上させるものである。比較する文に含まれている単語の一致する割合の計算は文脈ベクトル比較部７で行う。

【００４３】具体的には、図４のタイトルに含まれる単語は、「ノンバンク」、「整理」、「手続き」の３つであり、図４の本文の１番目の文に含まれるタイトルに含まれている単語は「ノンバンク」の１つ、２番目の文に含まれるタイトルに含まれている単語は「整理」、「手続き」の２つ、となっていて、ｉ番目の文のタイトルに含まれている単語と同一の単語を含んでいる割合ＳＴＫ（ｉ）は、

ＳＴＫ（１）＝１／３×１００＝３３

ＳＴＫ（２）＝２／３×１００＝６６

となる。同様にして、ｉ番目の文が、タイトルに含まれている単語と同一の単語を含んでいる割合ＳＴＫ（ｉ）を計算する。ｉ番目の文の文脈ベクトルとタイトルの文脈ベクトルとの内積値をＳＴＶ（ｉ）とすると、ｉ番目の文とタイトルとの距離ＳＴ（ｉ）は、

ＳＴ（ｉ）＝ｆｔ（ＳＴＫ（ｉ），ＳＴＶ（ｉ））

となる。ここでｆｔは、何らかの関数であり、単純に加算もしくはそれらの加重和を取る。具体例としては、

ＳＴ（ｉ）＝ＳＴＫ（ｉ）＋β×ＳＴＶ（ｉ）

となる。ここで、βは定数である。なお、β＝０の時は、タイトルに含まれている単語と同一の単語を含んでいる割合だけで要約文を選ぶことになり、β≫１の時はタイトルの文脈ベクトルとの内積値だけで要約文を選ぶことになる。

【００４４】また、ｉ番目の文が、文書全体に含まれている単語と同一の単語を含んでいる割合ＳＡＫ（ｉ）を計算する。ｉ番目の文の文脈ベクトルと文書全体の文脈ベクトルとの内積値をＳＡＶ（ｉ）とすると、ｉ番目の文のタイトルとの距離ＳＡ（ｉ）は、

ＳＡ（ｉ）＝ｆａ（ＳＡＫ（ｉ），ＳＡＶ（ｉ））

となる。ここでｆａは、何らかの関数であり、単純加算もしくはそれらの加重和を取る。具体例としては、

ＳＡ（ｉ）＝ＳＡＫ（ｉ）＋γ×ＳＡＶ（ｉ）

が考えられる。ここで、γは定数である。なお、γ＝０の時は、文書全体に含まれている単語と同一の単語を含んでいる割合だけで要約文を選ぶことになり、γ≫１の時は文書全体の文脈ベクトルとの内積値だけで要約文を選ぶことになる。

【００４５】以上のようにして求めたＳＴ（ｉ）とＳＡ（ｉ）とを、実施の形態２または実施の形態３と同様にして距離の順に、要約文が出力され、入力文書の要約となる。なお、βやγの設定は、文書出力部３の出力結果を見ながらユーザが文書入力部１から設定できるようにしてもよく、また、出力された要約文の所定順位までのＳＴ（ｉ）の総和とＳＡ（ｉ）の総和とが均等になるように文書処理部４で自動設定するようにしてもよい。

【００４６】（実施の形態５）本実施の形態は、実施の形態４において、タイトルに含まれている単語と同一の単語を含んでいる割合と文書全体に含まれている単語と同一の単語を含んでいる割合を計算する時に、単語辞書８に予め設定した単語の重要度を読み出して、含まれている単語の割合の計算に取り入れるというものである。

【００４７】上記ＳＴＫ（ｉ），ＳＡＫ（ｉ）は、

ＳＴＫ（ｉ）＝Σｐｉ／Σｐｔ×１００

ＳＡＫ（ｉ）＝Σｐｉ／Σｐａ×１００

と表すことができる。ここで、Σｐｉは文番号ｉの文に含まれている単語の単語の重要度の和を、Σｐｔはタイトルに含まれている単語の単語の重要度の和を、Σｐａは文書全体に含まれている単語の単語の重要度の和を表している。

【００４８】具体例を示してみると、図４のタイトルに含まれている単語の重要度を、「ノンバンク」を０．８、「整理」を０．５、「手続き」を０．６とすると、本文の１番目の文は「ノンバンク」の１つの単語を含んでいるので、

ＳＴＫ（１）＝０．８／（０．８＋０．５＋０．６）×１００

＝４２

となり、２番目の文は「整理」、「手続き」の２つの単語を含んでいるので、

ＳＴＫ（２）＝（０．５＋０．６）／（０．８＋０．５＋０．６）×１００

＝５８

となる。このように、重要度を取り入れることで、より精度の高い要約を得ることができる。

【００４９】以上の実施の形態においては、入力文書を文脈ベクトルを用いて解析することによって、従来の表層的な解析による上記第１の手法に比較して、文書における質の良い重要部分を抽出できる。また、従来の文章

を意味的に解析する第2の手法に比較して、事前に特定の文書形式や文脈を仮定する必要がないので、入力文書に対する自由度が大きく種々のタイプの文書に適用可能である。さらに、入力文書の構造解析や文脈の意味理解を行って内容の重要性を定義する必要がないので、より簡単な処理によって要約の抽出を行うことができる。また、上記第3の手法に比較しても、タイトルから要約に近い文脈ベクトルが構成されるため、より精度の高い要約文の抽出が可能である。

【0050】なお、上述した要約処理を実行するためのプログラムをフロッピーディスクやCDROM等のコンピュータ読み取り可能な記録媒体に予め記録させておいて、必要に応じてコンピュータにインストールさせて用いてもよい。

【0051】

【発明の効果】以上より明らかなように、本発明によれば、特定の文書形式や文脈を仮定することなく、簡単な処理にて意味的に重要な要約を高精度に生成できる。つまり、入力文書の中で不要な箇所を重要な箇所と誤ったり、仮定した文書形式や文脈に当て嵌まらない入力文書に対して全く非力であったりすることなく、種々のタイプの文書から適切な要約を抽出できる。

【図面の簡単な説明】

【図1】この発明の文書要約装置におけるブロック図である。

【図2】単語の文脈ベクトルが定義された単語辞書の内容の一例を示す図である。

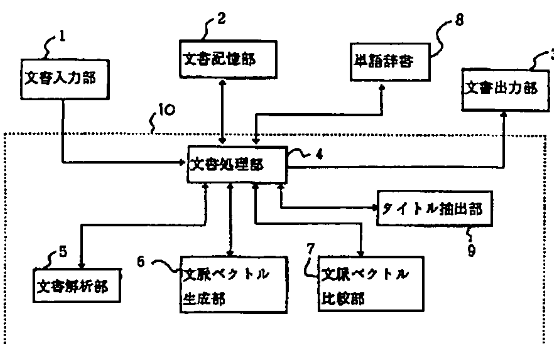
【図3】要約作成処理動作のフローチャートである。

【図4】要約作成処理動作を具体的に説明するための新聞記事の例である。

【符号の説明】

1 文書入力部
2 文書記憶部
3 文書出力部
4 文書処理部
5 文書解析部
6 文脈ベクトル生成部
7 文脈ベクトル比較部
8 単語辞書
9 タイトル抽出部

【図1】



【図2】

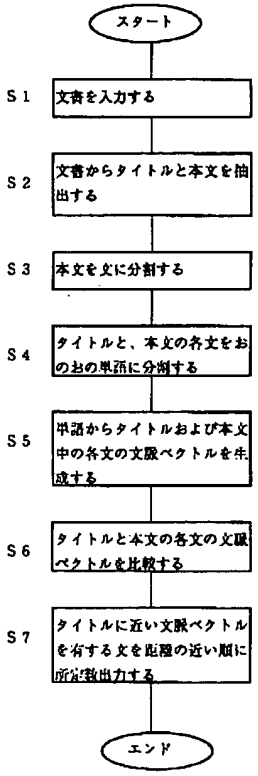
単語＼特徴語	人間 男 女 機械 知識 植物 経験 政治 芸術 科学
人間	1,1,1,0,0,0,0,0,0,0,.....
自動車	0,0,0,1,0,1,0,0,0,0,.....
読書	1,0,0,0,1,1,0,0,1,0,.....
薬品	1,0,0,0,1,0,0,0,0,1,.....
会社	1,0,0,0,0,1,1,0,0,0,.....
…	…
…	…

【図4】

タイトル： ノンバンク整理手続き

本文： 日本債権信用銀行の関連ノンバンク三社は、銀行などへの元利支払いを停止した。一日にも不良債権処理のため整理手続きに入り、事実上倒産する。整理の方法は特別清算が有力。損失補填は日債銀も他の融資金融機関と同様に融資残高に応じ負担する比例配分方式となる見込み。これを受け、日債銀は海外撤退を・・・・・・・・

【図３】


スタート
S1
文書を入力する
S2
文書からタイトルと本文を抽出する
S3
本文を文に分割する
S4
タイトルと、本文の各文をおのおの単語に分割する
S5
単語からタイトルおよび本文中の各文の文脈ベクトルを生成する
S6
タイトルと本文の各文の文脈ベクトルを比較する
S7
タイトルに近い文脈ベクトルを有する文を距離の近い順に所定数出力する
エンド